penalty. This is the reason for certain cut-offs for intrachain distances and dot products of the relevant side-chain vectors. Of course, these cut-offs are consistent with the vast majority of helical or β-type geometries seen in globular proteins. Of course, these terms may be modified or refined as additional three-dimensional proteins structures are solved to high resolution

Long-range constraints

Long-range *constraints* are implemented in the form of a distorted harmonic potential. Additionally, the contact energy for such side chain pairs is modified as well below:

$$E_{ij,\ \text{restrained}} = \begin{cases} \infty, & \text{for } r_{ij} < 3 \\ E^{rep}, & \text{for } 3 \le r_{ij} < R_{i,j}^{rep} \\ \in_{i,j} - 0.5, & \text{for } R_{i,j}^{rep} \le r_{ij} < R_{i,j} \\ \in_{res} (R_{i,j}^2 - R_{i,j}^{2\,rep} & \text{for } R_{i,j} < r_{ij} < 10 \\ \in_{res} (100 - r_{ij}^2 - 100)/3 & \text{for } 10 < r_{ij} \end{cases} \quad (14)$$

The value of parameter $\in_{res}$ in structure assembly runs was set equal to 1/8, while during the low temperature refinement run, it was set equal to ¼. The meaning of other parameters is the same as in equation 8, above. In the first three ranges, the above function is consistent with the definition of pairwise interactions defined in the previous section. For restrained residues, the pairwise potential has been enhanced (line 3 of equation 14). The two remaining lines define a pseudoharmonic long-distance potential. For longer distances (line 5 of equation 14), it is slightly suppressed because a weaker function facilitates a somewhat faster assembly of model protein chains.

Folding Procedure

The sampling procedure employed for protein assembly is based on Monte Carlo simulated thermal annealing. The stages are described below:

1. In the first step, a random expanded chain conformation is subjected to Monte Carlo simulated thermal annealing[38] over a broad range of temperature from T = 6 (T = 4 for smaller proteins) toT = 1. After annealing, the number of satisfied long-range constraints in each folded protein is inspected. Those folds with more than about 1.7 of their constraints significantly violated are rejected without further inspection, *e.g.*, when the corresponding side-chain:side chain distance is larger than 7 lattice units for proteins smaller than about 100 residues and 8 lattice units for proteins larger than about 100 residues. These alternative, exemplary parameters have been selected by studying a similar problem,[6] and by preliminary testing of the present model. Allowing a significantly larger number of violated constraints may lead to topologically wrong folds, while requesting all constraints to be satisfied would decrease the efficiency of the method, as some good folds with small local distortions would be rejected. The success ratio at this stage depends on the protein and the number of long-range constraints. For example, in 1gb1, protein G, with eight constraints, more than 75% of short assembly runs (5-15 minutes of CPU time on a HP C-110 workstation) are successful. In the case of 1pcy, plastocyanin, with 15 constraints, the corresponding success rate is about 30% for 4-hour-long simulations on an HP C-110 workstation. Of course, a slower annealing protocol increases the fraction of assembled structures that satisfy the constraints. However, it appears that use of a larger number of shorter simulations is a more effective sampling protocol because a greater number of structures are collected for each protein.

2. All structures obtained via the rapid annealing procedure are preferably subjected to a refinement process. For refinement, each structure is duplicated and subjected to two independent Monte Carlo annealing runs over the

temperature range T = 2-1. The lowest conformational energy structure (from the last snapshot of the corresponding trajectories) is accepted for further analysis.

3. For each protein, both the lowest energy conformation and the lowest energy alternative conformation are then subjected to isothermal runs to establish whether the proper fold can be automatically selected based on the choice of the lowest average energy structure.

4. The Cα coordinates of the final, lowest energy structures are then built into the model. This "back filling" procedure is based on Monte Carlo annealing of a phantom lattice model chain that has two united atoms per residue: one centered on the Cα and the other at the side chain center of mass. This Cα plus side chain center of mass, CAPLUS, model (*see, e.g.*, [6,16,29-31,39]) only employs statistical potentials describing short-range interactions and side chain rotamer preferences. The positions of the side chains in the CAPLUS model are driven by a harmonic potential to the predicted side chain positions from the side chain only model.

As those in the art will appreciate, other models of differing levels of detail, up to an including all heavy atom, and even all atom, representations, can also be assembled from these low energy structures. The level of detail and resolution chosen for these structures will typically depend on the particular application for which the model is intended. For example, rational drug design typically requires models having significant levels of atomic detail, particularly when protein:ligand interactions are being assessed.

APPLICATIONS AND AUTOMATED IMPLEMENTATION

Protein Function Determination

As described above, it is now possible to rapidly generate accurate, reduced protein models directly from nucleotide or deduced amino acid sequence data. These models, which are based on the side chain center of mass of the amino acid